USING DIRECTIONAL ASYMMETRY TO ASSESS SELECTION IN COPY NUMBER VARAINTS

Benjamin D. Greenbaum[1], Chang Chan[1], Asad Naqvi[1], Lane McIntosh[2], Arnold J. Levine [1]

[1]Simons Center for Systems Biology, Institute for Advanced Study, Princeton, New Jersey, [2]Department of Mathematics, University of Hawaii, Honolulu, HI.

Platform independent

Vs disease

The contribution of Copy Number Variations (CNVs) to overall human genetic variation has become clear only recently, with estimates of its extent that would have been surprising a decade ago []. Many CNVs overlap with genes, highlighting their potential role in evolution and genetic disorders []. While many studies focused on the potential role of CNVs in disease, putatively positive effects, such as protein dosage change or increased robustness against deleterious mutations, have also been noted []. For example, duplications of the salivary starch digestion enzyme alpha-amylase (*AMY1*) have been linked to salivary concentrations of this enzyme in populations with starch-rich diets []. Similar hypotheses of advantageous CNV effects exist for other genes, such as the *CCL3* locus or a set of beta-defensins [].

Both the number of genes subject to positive effects and the method for assessing and quantifying these effects remain unclear []. Researchers initially hypothesized that many detected CNVs underwent positive selection because CNVs occurred disproportionately in "environmental genes," such as those involved in defense and metabolism []. This view is no longer generally accepted, with many researchers adopting the belief that the disproportionate amount of environmental genes amongst common CNVs reflects the presence of many of these environmental genes at variational hotspots, such as segmental duplications (SDs, also known as low-copy repeats), combined with reduced purifying selection of environmental genes compared to essential genes []. The combination of reduced purifying selection and high variability causes these genes to appear more frequently in a population than essential genes, which are less tolerant of and, often, less prone to variation. Thus the frequency of their presence can be statistically significant in population studies that examine functional categories of genes prone to copy changes [].

Measures for selection on CNVs have typically either focused on frequency and functional categories of heavily duplicated or deleted genes, or have modified methods for individual polymorphic sites and single nucleotide polymorphism (SNP) based haplotype linkage methods. These have disadvantages. Gene duplication could offer a dose advantage without altering a single internal amino acid, confounding mutational methods []. If, for instance, in one lineage, a gene acquired a CNV that has a dosage advantage, while the non-CNV lineage acquired a beneficial point mutation, this could be

difficult to deconvolve. While CNV and SNP information may turn out to be redundant in the long run, as suggested by Ref. [], ideally this would come from the convergence of methods designed for each task, rather than a reliance on SNP approaches to characterize positive selection in CNVs. Likewise, methods that have tended to focus on frequency and functional categories of heavily duplicated or deleted genes have substantial drawbacks, since many of these genes may be part of the same CNV and may have been varied together, invalidating the p-value methods used to establish functional significance []. This paper introduces a novel, self-consistent, quantitative approach specifically designed to evaluate directional selection on CNVs, without relying on methods previously designed for other types of variation and taking advantage of previous experimental work on the empirically-derived statistical properties of this type of variation. The most common CNV-producing mechanisms, such as interchromosomal non-allelic homologous recombination, are symmetric processes, which introduce a duplication and deletion simultaneously into a population while preserving the overall number of genes []. If an intrinsically variable region produces both duplication and deletions, yet one is far more prevalent than the other, this may signal that the CNV causes an advantageous trait. Therefore, using a well-conceived statistical measure of bias for duplications versus deletions at a given loci will provide criteria for selection that do not depend on other proxies for its support.

This paper proposes such a measure. Below, we outline our method for detecting selection on CNVs, which separately quantifies outliers for absolute occurrence relative to the population (addressing the issue of reduced purifying selection) and site-specific directional bias (addressing whether or not the previously defined outliers are directional relative to their site specific variation). We effectively separate genes by three successive filters of CNV occurrence, with the first being random noise fluctuations in CNV direction. The second narrows the search to total absolute event counts outliers for the whole population, as quantified by a negative binomial distribution. Due to this distribution's ability to capture heterogeneous counting processes, it can represent the effects of reduced purifying selection, building this confounding effect into the test. Finally, we quantify outliers of the previous group by whether they are strongly directional given the variability of that site. Therefore, the surviving, small set of genes have both a large occurrence of gene-overlapping CNV events relative to the population's rate of CNV event occurrence and a site specific directional bias, which accounts for the probability of the observed directionality given the individual site's variability.

This method is tested in several large datasets, and the candidate genes for positive selection are found to be invariant across these sets and strongly overlap with segmental duplications. This group contains three large subsets of genes involved in (1) the metabolism of digesting and processing starch and xenobiotic intermediates, (2) defense and immunity, and (3) olfaction. This is a subset of the environmental gene categories previously observed in CNV processes as being high frequency, but many more environmental genes previously proposed in the literature do not survive these statistical filters. Many of the genes that survive have previously been suspected of having phenotypic dosage effects (*AMYL*, *HLADRB*, *DEFB*, and *FGCR3* loci), providing additional proof of concept for this method []. More strongly, some identified genes,

particularly those involved in xenobiotic metabolism, come from independent loci in the genome yet participate in similar pathways. Finally, we employ haplotype-based methods to provide validation for these selection procedures in the cases where these regions overlap with linked haplotype blocks. We conclude that this new approach identifies a set of CNVs best explained by positive selection.

RESULTS

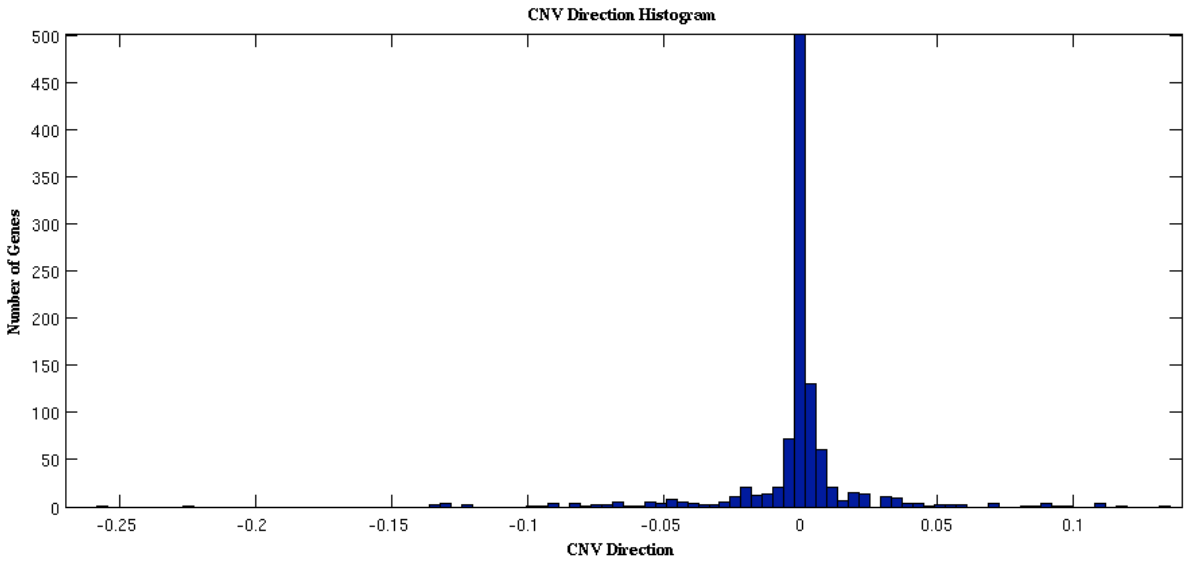
COPY NUMBER HAS A STRONG DIRECTIONAL BIAS IN A SUBSET OF GENES

A set of 1259 unrelated Caucasians used as the control group in a schizophrenia study (referred to as Dataset 1), performed on the Affymetrix 6.0 SNP array platform, were initially examined []. A CNV is assigned whenever more than three consecutive probes consistently indicate either duplication or deletion, the idea being that one should initially have a weak criteria for calling an event so that the noisy regime will be more easily defined. Each CNV is given a score from [-2:2]: a deletion on both chromosomes counts as -2 and on one as -1, while duplication on one chromosome counts as 1 and on both chromosomes as 2. For each RefSeq gene, the *CNV direction* of a gene, is the sum of the scores for all CNVs that overlap with a given gene, divided by the total number of autosomes sampled in the population (in this case 2518). Assuming a gene overlaps with T CNV events and the CNV score for each overlap is labeled s_k, the CNV direction, using the notation d, of a gene is defined as

$$\frac{1}{2N}\sum_{k=1}^{T} s_k$$

where N is the total number of individuals. Only autosomes were employed for this analysis to avoid complications. Figure 1a is a histogram of CNV directions for all annotated RefSeq genes:

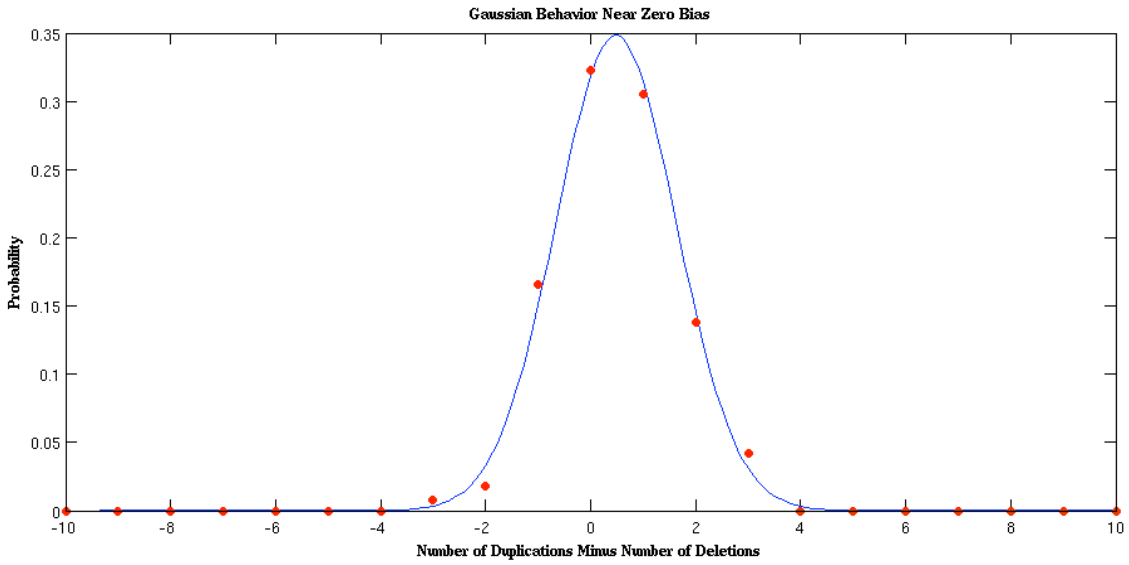
1a)



Most genes are not involved in any CNV events, or, if involved, lack a strong preference for duplications or deletions. They appear as fluctuations, likely from measurement noise and real low frequency events. This is fit by the maximum likelihood best-fit normal distribution with the same mean and standard deviation as the data on various symmetric intervals about the x-axis. This procedure also varies the zero count, from the largest recorded count to all uncounted RefSeq genes, which is the greatest possible value. The zero level is varied because it is unclear how many unrecorded genes could have been involved in CNVs. The Kullback-Leibler divergence is then used to find which best normal fit is closest to the empirical distribution, both the zero estimate.

The resulting normal distribution is supported on the interval of [-3:3] as pictured in Figure 1b ([-0.0012:0.0012] autosomes). Its mean is not zero, but 0.48, possibly because a deletion is more likely to cause loss of function, resulting in stronger elimination pressure (as also noted in Ref. []) or because of unaccounted multiple duplication events. The former interpretation is supported by the 2.26 ratio (31221 vs. 13793) of deletions over duplications. The total zero estimate for this region is 1822 genes, involving a total of 5651 genes between [-3:3] out of 19655 RefSeq autosomal genes. The standard deviation is a narrow 1.15.

1b)

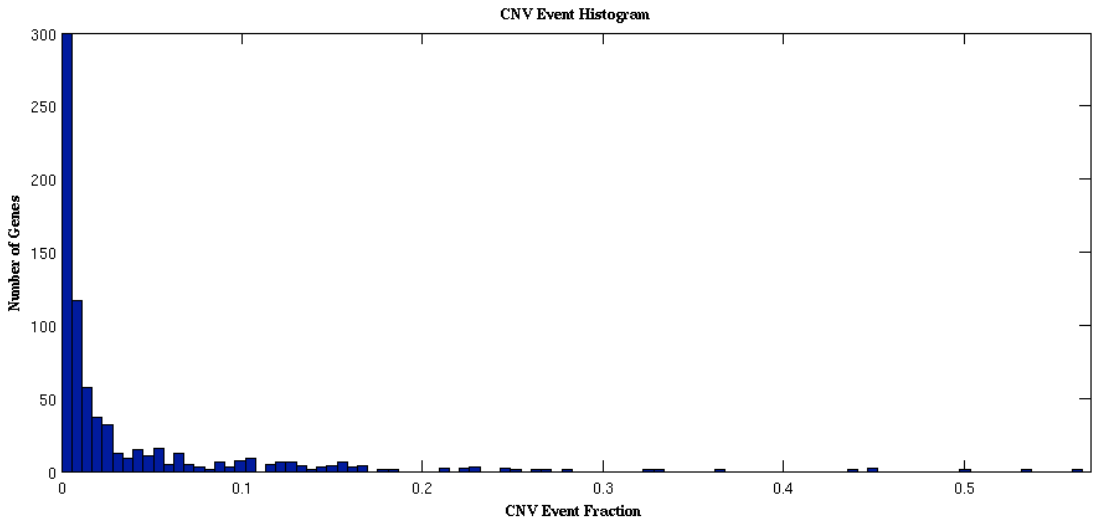


Consistent with the above observation (along with the mechanisms of Refs. []) that the random part of the distribution is virtually directionless, though slightly skewed towards duplications, each duplication or deletion that overlaps with a gene is counted as an "event". The total number of events per gene, using the notation *n*, is defined as

$$\frac{1}{2N}\sum_{k=1}^{T}|s_k|$$

Figure 2a shows the event distribution for all genes, with the inferred zero value:

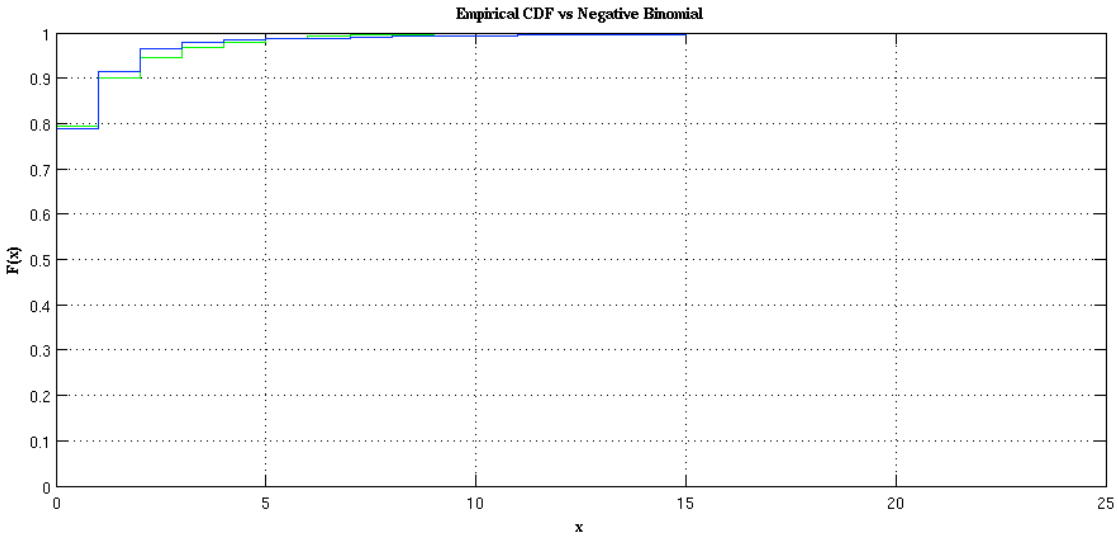
2a)



This counting process is *highly* over-dispersed (i.e. the variance is much larger than the mean), indicating a negative binomial distribution is a candidate for fitting []. This distribution captures a gamma distributed set of Poisson processes, often a good model for heterogeneous aggregation or contagious processes []. The analogy here is with reduced purifying selection: as some groups of genes within CNV generating regions are more slowly eliminated and/or come from high variability regions, their Poisson event rate in the population will be larger than the event rate for more rapidly eliminated and/or less variable genes, leading to a heterogeneous set of event rates and aggregated counts. The distribution generally describes the overall probability of j events occurring with probability p when r failures are tolerated. In the preceding interpretation, r and p, parameters to be fit, dictate the shape of the gamma distribution of Poisson events rates, with the gamma distribution's parameters becoming r and $p/(1-p)$.

For each j value, we calculate a maximum likelihood negative binomial fit to all data below j events and compare this to the corresponding empirical cumulative distribution using the two-sample Kolmogorov-Smirnov test. The p-value for agreement is below 0.05 after a total of 19 events, meaning one can expect to see up to 19 duplications or deletions per gene under this model. Even inflating to the largest possible zero value (the total number of RefSeq genes) would only increase the cutoff from 19 to 21, since the negative binomial is not especially sensitive to its zero-value. For 19 events, the negative binomial describes the empirical cumulative distribution well, as illustrated in Figure 2b:

2b)



Since only 3 consecutive SNPs were used to call a CNV, if outlier genes always contain few SNPs, weak calls may be responsible. In Dataset 1, the median number of SNPs that form a CNV event is 30, with a skewed mean of 48.2, and the median length of a CNV event is 25.4 kb, regardless of whether that event overlaps with a gene (an estimate given heterogeneous probe spacing). For duplication-prone genes, those who are outliers in both direction and absolute number, the median and mean number of SNPs involved in a CNV event overlapping with a gene is 77 and 127.6, respectively. This may also come from the fact that, as duplications are intrinsically more difficult to call, low SNP duplication calls are difficult to make. For deletion-prone genes, these values are 34 and 52.7 respectively. The apparent difference between these two groups is reinforced as duplication-prone genes are easily distinguished from the overall distribution of SNPs comprising a CNV by the Mann-Whitney statistical test ($p<0.0001$), whereas the deletion-prone genes are not ($p=0.63$). Therefore, while deletion-prone genes have a CNV profile more like a typical CNV, duplication-prone genes are more robust. This assuages any doubts that these effects come form weak calls.

As shorter genes may be more easily copied, particularly as a whole, one may expect gene length to play a role in creating these outliers. In fact, the median length of duplication-prone genes is 13.6 kb while this value is 35.5 kb for deletion-prone genes. The undirected but variable genes more closely resemble the duplication-prone genes, with a median length of 14.6 kb. For a gene involved in any CNV event, the mean fraction of that gene's length overlapping with the CNV is 0.63, with 1 as the median value. For highly duplicated genes, 0.847 coverage is the mean value (1 is the median). However, for a deletion to cause loss of function, only part of the gene may need to be deleted. Clearly genes smaller than the typical CNV size are amongst the duplication-prone, implying their length is often preserved, and are also like the generally variable genes. As hypothesized in the literature, segmental duplications may generate CNVs, and this seems to be true amongst these duplication-prone genes []. Using the UCSC database of segmental duplications, we tested if the duplication-prone set is more likely than not to overlap with a segmental duplication []. While only 2853 of the 19655

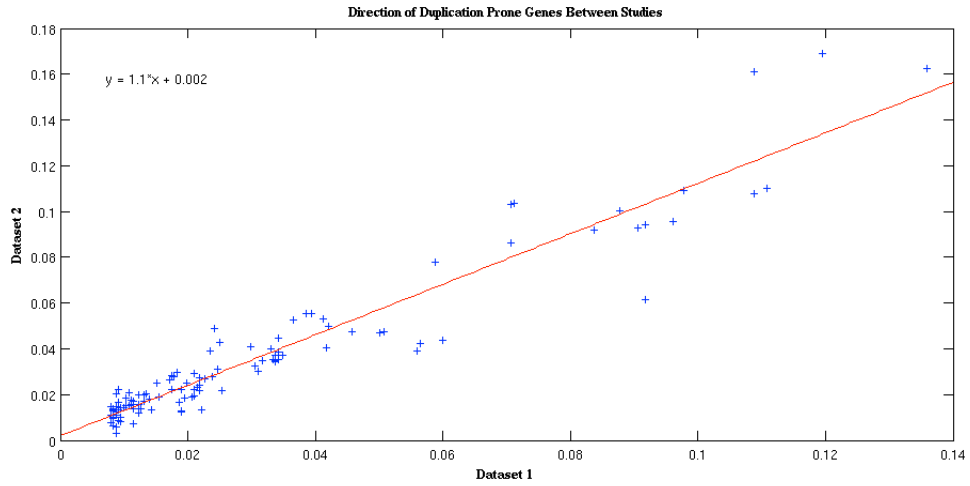
RefSeq autosomal genes overlap with an SD, 78 of the 134 duplication-prone genes overlap, indicating that duplication-prone genes strongly overlap with low copy repeat regions (Fisher's Exact p-value=6.56e-19).

As a final observation, a CHOP dataset of 2026 individuals was examined, containing 1335 Caucasian and 697 African-American participants []. These data are obtained from an Illumina 550 Chip with lower coverage, a chip that does not discriminate between one or two duplications, but does discriminate between one and two deletions. Therefore, our directional measure must be modified to count CNVs on both chromosomes and one chromosome equally, using values [-1:1]. If restricted to the Caucasian subset of these data, due to the smaller sample size and lower coverage, one detects 22 duplication-prone genes, 19 of which overlap with the previous duplication-prone genes. For deletion-prone genes, there are 78 genes, only 29 of which intersect with the consensus set genes, likely reflecting again that deletion outliers are more difficult to interpret. The above serves as remarkable reinforcement of the observed CNV bias on a different, noisier chip, particularly for duplication-prone genes. For the African-American population, there are clear differences. Due to the smaller sample size, only 15 duplication-prone genes are deemed significant and only 3 of them overlap with the previously found set, while only 11 of the 60 deletion-prone genes overlap. This reinforces earlier, limited studies, which found differences between African-American and Caucasian CNVs with similar platforms and population sizes [].

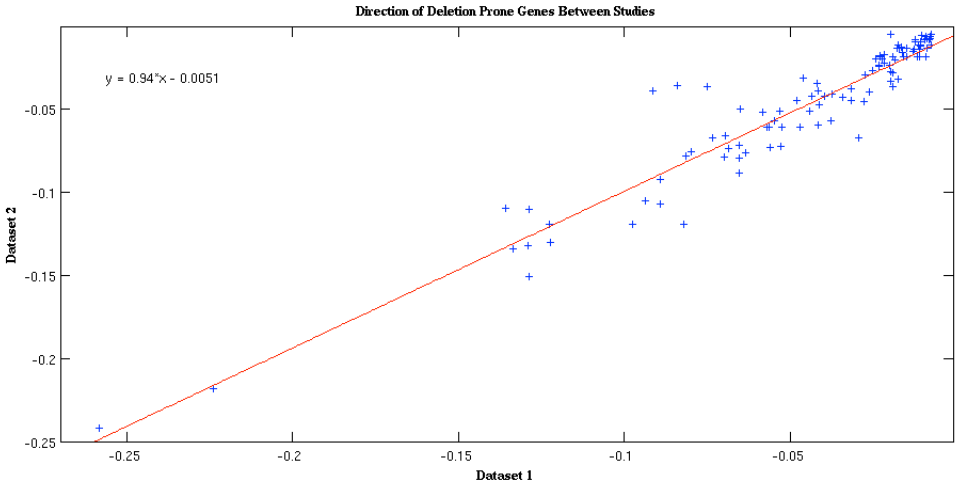
DIRECTIONAL BIAS

To determine if these results are robust across studies, the analysis was repeated for a diabetes study, "Dataset 2," consisting of 3032 primarily Caucasian cases and controls, on the same Affymetrix chip []. No significant differences are found between the diabetes case and controls, allowing us to use all study members in deriving future statistics. The two sets also have similar directional values, as seen in Figure 3a and b where the value of the CNV direction for the most directionally biased genes in both studies is compared, and found highly correlated. As a consequence, to assess the significance of genes with the strongest directional bias, both datasets are concatenated.

3a)



3b)



Applying the negative binomial test to this group yield 601 genes with an event rate that is higher than the observed event rate distribution in the total population from the combined datasets. As observed above, the outliers of the negative binomial set have many members that are also outliers in their CNV direction relative to the population. However, this may simply be due to the fact that these CNVs are more common, and the directional bias may occur by chance. This can be tested directly, by quantifying the direction, given the site's total number of events. This makes each site a binomial random walk for the direction given the total variability. Thus the two-tail binomial probability for the CNV direction, *d*, of a gene, given the total number of events for that gene, *n*, is

$$\frac{1}{2^{n-1}} \sum_{i=(n+d)/2}^{n} \binom{n}{i}.$$

By applying this measure to the 601 outlier genes, and using a p-value cutoff of 0.05/601 to correct for multiple hypothesis testing, we get 411 genes that have a population wide high event rate *and* are significantly directional for their site. If we look at those with a positive direction, this yields 233 duplication prone genes, and those with a negative direction are 178 deletion prone genes. There are 190 genes that are eliminated by this method, and interestingly, only 20 that are highly directional but miss the negative binomial cutoff, essentially estimating the false directional positives and negatives if one were to use the negative binomial approach alone. The highly directional genes, with their total event count and direction are listed in Supplementary Table 1. Hence our final method uses a two step filtering process: the negative binomial with adjusted zero (though in practice this value has little effect on the results) to obtain a set of genes whose event rates stand out strongly from the population spread of event rates, followed by the binomial direction probability on those outlier genes given their site variability.

FUNCTIONAL CATEGORIES AND POSSIBLE SELECTIVE TRAITS

The validated genes were explored for their functional categories. Studies have focused on function to imply positive selection, and screening the list of genes through standard tools such as Gene Ontology, KEGG and Ingenuity indicates that the most represented categories amongst duplication-prone genes are olfaction, immunity and metabolism (Supplementary Table 2), while there is consistently less significance assigned to the categories of deletion-prone genes (unpublished observation). However, these methods do not capture the full picture because they assign p-values to each gene as if they are independent, even though CNVs can disrupt or copy multiple genes simultaneously. We therefore group any genes that appear together in CNV events to form clusters of co-appearing genes. Selection may preserve these groups because they contain one or more genes that offer a benefit and the other genes detected are hitchhikers, preserved because they lie on a frequently duplicated block with a beneficial gene. While curated databases pick up defense, metabolic and olfactory genes as being the most significant groups, their effect may be over- or underestimated as they can cause duplicated blocks to carry genes unassociated with these categories. Of the 233 consistently duplication prone genes, there are only 81 independent clusters by this method (Supplementary Table 3), and thus many genes counted by looking at, say Gene Ontology categories, are not strictly valid. For deletion prone genes, this clustering occurs far less frequently (Supplementary Table 4).

Moreover, in each of these clusters, there is often a gene involved in a functional category appearing with an unannotated gene. For instance, 42 of the 81 duplication-prone clusters contain one or more genes involved in immunity or defense, metabolism, or olfaction, so that one may suspect they are driving the process, accounting for 122 of the 233 duplication prone genes. In fact one of those genes is usually the most frequently duplicated, confirming that suspicion. Initially the categories involving metabolic processes are examined. The main clusters of these genes are pictured in Table 1, and the most frequently noted gene is italicized:

Chromosome	Start	Stop	Genes
1	9217449	9352175	*H6PD* *SPSB1*
1	103898844	104102834	*AMY1A* *AMY1B* *AMY1C* AMY2A AMY2B LOC648740
1	110012166	110037889	*GSTM1* GSTM2
2	38746555	38832140	*GALM* SFRS7
4	68995761	69570979	TMPRSS11E TMPRSS11E2 UGT2B15 *UGT2B17*
10	135117421	135232866	*CYP2E1* LOC619207 SYCE1
11	107166928	107235124	SLC35F2
12	7857663	7980159	FAM90A1 SLC2A14 *SLC2A3*
16	16333234	18517108	*ABCC6P1* LOC339047 NOMO2
16	32070109	33172228	HERC2P4 *LOC729355* SLC6A10P *TP53TG3*
16	54394264	54424576	CES1
22	22670594	22677258	GSTTP1

The first category is starch and sucrose metabolism, confirming the finding of the amylase studies, as five alpha-amylase genes are duplication prone. While Ref. [] focuses on the salivary *AMY1* genes, we find that the pancreatic *AMY2* genes are also highly duplicated. The KEGG database groups *UGT2B15* and *UGT2B17* in this pathway, but they are more associated with the glucouronidation reaction downstream of amylase itself. This leads to the most interesting finding of a seven-gene set associated with xenobiotic metabolism across 4 clusters on separate chromosomes. These genes are *CYP2E1*, *GSTM1*, *GSTM2*, *GSTTP1*, and the previously mentioned *UGT2B15* and *17*. While *GSTTP1* is a pseudogene, it is nearby functional homologs that barely missed the cutoff. The first is a phase I xenobiotic component which introduces a reactive group to the xenobiotic and is most associated with alcohol and drug detoxification []. Next we find the phase II proteins such as *GSTM1* and *GSTM2* glutathione S-transferases that attach charged groups to reactive sites, and *UGTB15* and *17*, which add sugars to polar xenobiotics so that they are more easily excreted. The fact that multiple loci contribute additional copies to a process involving many genes where quantity dependence has a clear benefit makes a persuasive case that selection increased the efficacy of xenobiotic metabolism. Interestingly the glucuronosyltransferase genes implicated in both metabolic pathways have the highest percentage of duplications (one is heterozygous in 38% of the population). As this protein is generically involved in many metabolic processes, there might be a greater selective pressure.

The primary immune and barrier defense and maintenance clusters are in Table 2:

Chromosome	Start	Stop	Genes
1	159817761	159914575	*FCGR2C* HSPA7
1	151035850	151037281	LCE1D
3	101811134	101950501	*GPR128* TFG
3	120102168	120347588	IGSF11

3	196933423	197023545	*MUC20* MUC4
4	8978697	9095260	*DEFB131* DUB4 LOC650293
6	32593131	32665540	HLA-DRB1 *HLA-DRB5* HLA-DRB6
7	142979011	143708658	ARHGEF5 *CTAGE6* FAM115C *LOC154761*
			LOC441294
8	7742811	8139796	*DEFB4* FLJ10661 SPAG11A
8	11959306	12338223	DEFB109P1 DEFB130 FAM86B1 *FAM86B2*
10	46077145	51563275	ANTXRL ANXA8 ANXA8L1 ANXA8L2
			BMS1P1 BMS1P5 FAM21B *FAM25B*
			FAM25C FAM25G FRMPD2L1 FRMPD2L2
			GPRIN2 LOC642826 LOC728643 PPYR1
			SYT15
12	50981915	51065684	KRT83 KRT84 KRT85 *KRT86*
12	10889714	11215480	*PRH1 PRR4*
17	31517173	31882204	*CCL3L1 CCL3L3* CCL4L1 CCL4L2 TBC1D3B TBC1D3C
19	59432280	59956316	KIR2DL3 KIR3DL3 *LILRA6*
20	1493029	1548689	SIRPB1
21	10042712	10120808	BAGE *BAGE2 BAGE3 BAGE4 BAGE5*
22	24043887	24107544	IGLL3 *LRP5L*

The beta-defensins and *CCL3*, found in previous studies, are recovered here, along with a class two HLA region and genes involved in antigen presentation []. Other interesting defense genes include annexins, involved in inflammation control, mucin, which protects epithelial surfaces, and *PRR4*, which protects the surface of the eye. Additionally one finds several immunoglobulin-related genes on multiple chromosomes, *IGSF11*, *SIRPB1*, and *IGLL3*. The *FGCR* locus, also implicated in a previous study, was recovered [].

The olfactory clusters are listed in Table 3:

	Start	**Stop**	**Genes**
1	357521	611897	*OR4F16 OR4F29 OR4F3*
11	4923964	4924906	OR51A4
11	55127492	55128425	OR4C11
14	18447593	19474601	OR11H12 OR4K1 OR4K2 OR4K5 OR4M1 OR4N2 *OR4Q3* P704P POTEG
15	18997107	19915749	CXADRP2 GOLGA8C LOC646214 LOC650137 LOC727832 *LOC727924* OR4M2 POTEB
15	100163445	100176851	*OR4F15* OR4F6

As a final filter for the number of functional categories, we explore the possibility that a set of genes were generated by an asymmetric, less frequent process that favors duplications over deletions []. One way to test this is to tag clusters to see if they only contain duplication-prone genes, and never a deletion. Most clusters eliminated by this

method are not involved in the above processes and correspond to regions consisting of RNA that is not well characterized, such as cluster 55, which only contains highly repetitive small nuclear RNA (such clusters are indicated in Column 1 of Supplementary Table 3 by an asterisk). This method applies to 10 of the 81 clusters, making the functionally significant clusters more robust. Ideally, one would like to identify something more specific about these regions and eliminate these blocks to provide an additional final filter.

While deletions lack strongly significant categories, (save the presence of olfactory receptors) there are still interesting deletion-prone genes. One finds the diabetes-implicated *AKT3*, the cancer suppressor *TUSC3* or the ErbB pathway genes *ERBB4* and *MAPK10*, reinforcing the importance of significant CNVs to disease study. Intriguingly, a set of metabolic genes are also present here, some of which seem to complement the duplications. *GSTT1* is present, which is a less common glutathione S-transferase than *GSTM1*. Likewise, *MGAM*, maltase glucoamalyse, a less common starch digestion enzyme, is also deletion-prone. These genes are in highly variable regions and, as these genes become less utilized due to compensation by duplication, their deletions may be very slowly eliminated.

HAPLOTYPE AND POPULATION GENETIC VALIDATIONS

If one had a large set of reliable sequences, and knew which CNVs were functional in providing gain or loss of dosage, then a population genetics study of CNVs would be more feasible. For deletions, it does not seem possible to investigate further without experimental verification of which deletions alter phenotype, given that deletions often do not cover whole genes. A fragment of a gene being deleted may be meaningless, or may indicate a complete loss of function. For a duplication to have an adaptive, dosage-increasing effect, a whole copy should be duplicated. Even this can be difficult to detect: if chip coverage stops before a gene boundary and does not resume until far from that site, it will be inconclusive if a whole gene has been copied. Cases where a whole gene copy is inconclusive are listed in Supplementary Table 5 as NA. The other columns represent the homo- and heterozygosity for the trait of having a whole gene copied, while the third column is the chi-squared test for the first two columns. There are two key observations. The first is that, for a handful of genes, Hardy-Weinberg equilibrium for the trait of having the whole gene copied actually seems to be a decent assumption. Moreover, there is wide variation in chi-squared values, indicating that either they are in the categories or there is real disequilibrium.

Another approach is to employ haplotype-based methods for positive selection []. The applicability of these methods for CNVs has been discussed in the literature in some detail, but here we only use these approaches for validation. A region is examined for linked SNPs to see if haplotype blocks are within 20 kb of the transcription start or stop sites of a validated gene. Because some of these genes are nearby each other, only unique regions are used. Only 25% of the duplication prone genes have a haplotype block in their vicinity. The population is then divided into cases and controls for duplications (or deletions) of that gene. To gain statistical significance both homo- and

heterozygotes are counted as cases. While this avoids haplotype phasing issues, it can add additional noise.

In the blocks overlapping with duplication-prone genes, there is a discrepancy in terms of both diversity (defined as in Ref. []) and length between the two datasets. To test the significance of this disparity we performed a permutation test for these two groups, randomly resampling their categories 10000 times. We performed this separately for both datasets. In both sets (Table 4) there was low diversity in both case and control and a significant difference in diversity between case and control, with the cases having lower diversity. The average haplotype block length was longer for cases in both sets, though this difference was less significant. We performed the test for those genes that were highly variable, from the negative binomial cutoff, yet missed the directional cutoff. These genes, referred to as the remainder set in Table 4, have higher diversity and a less significant case-control difference. They are a significant weakening of the observed effect in the duplication-prone genes. Intriguingly, deletion-prone genes, when individuals with deletions are removed from the population, show a much higher overall diversity (0.97), reinforcing the significance of the previous results, as well as the observation that deletion prone genes are less likely to indicate selection.

	Dataset 1		**Dataset 2**	
	Case	*Control*	*Case*	*Control*
Duplication Prone				
Number of Blocks	228	380	373	466
Number of Genes	58	58	86	86
Average Diversity	0.5525	0.5869	0.5714	0.5908
Diversity P-Value	0.0161		0.0452	
Average Length	20.22	16.56	19.21	16.73
Remainder Set				
Number of Blocks	372	495	317	621
Number of Genes	20	20	40	40
Average Diversity	0.6594	0.6889	0.6176	0.6359
Diversity P-Value	0.0169		0.069	
Average Length	11.75	12.54	15.51	19.72

CONCLUSION

In this paper, we introduce a new method for detecting positive selection based on copy number variation that is tailored to the underlying variation generating processes. This novel approach, based on statistically stratifying genes by both CNV magnitude and direction, identifies genes with significant overall rates of appearance in CNV events and directions given the number of CNV event appearances. Gene directedness is similar

across several datasets. Duplication-prone genes are relatively short and are typically copied as a whole, thus providing a possible dosage benefit. Deletion-prone genes, on the other hand, are typically only copied in part, and are less likely to indicate selection. Annotation may be at fault. For instance, *TP63* is deletion-prone because it has multiple splice variants and a 1 kb block missing from its longest intron. There are no known functional consequences described in the literature for this TP63 deletion, and it does not replicate the phenotype of missense mutations. It is difficult to assign meaning to deletion outliers, such as TP63, without clear references and without knowing exactly which are decreasing dosage due to lack of function.

The case for positive selection is much stronger for duplication-prone genes. Many duplication-prone genes are copied as a whole and are associated with longer, less diverse haplotypes. Moreover, duplication-prone genes form clusters with representative genes from functional groups. A majority of genes are either in the functional groups olfaction, immunity, or metabolism, or are close to a gene in one of these groups. Within immunity and metabolism, there are specific subcategories to which gene groups on multiple chromosomes contribute. In the case of immunity, there are several gene groups involved in innate signaling and intrinsic defense; whereas the metabolic genes favor starch processing and xenobiotic metabolism. In this sense, our method is predictive, as each block that has a statistical and functional association for gene duplication should have an advantageous effect upon at least one gene in that block.

A typical duplication-prone gene is within a set of genes in a block, with one or more from the same functional category. This configuration may significantly impact how CNVs affect disease. Not only do CNVs tend to create clusters of homologous genes, which can then diversify through mutation, but this study also shows that beneficial genes duplicate in a block along with other genes that may be unrelated in function. These unrelated genes could be involved in a disease or present unintended consequences. For instance, *CTAGE6*, a T-cell associated antigen, often duplicates along with a set of nearby olfactory receptors. A beneficial gene that is selected for and carries along one that can cause a disease under other circumstances (exposure to an allergen or changes in the environment) could fix the non-selected gene into the population at a high frequency. Interestingly, some genes identified here were recently associated with autoimmune disease in Ref. [], suggesting that what was beneficial to previous generations in fighting infection may leave one prone to autoimmune disorders in a new or different environment, while at the same time validating that these genes likely have a phenotypic dosage effect.

This paper presents a method for identifying genes that are subject to copy number variation and are likely to contribute advantageous effects. More broadly, the novel methodology presented here identifies selection on a level where the variational background is separate from point mutation. Understanding this and other mechanisms at the different levels on which selection occurs is part of an overall effort to complete our knowledge of what genetic alterations alter phenotype. In this era of high volume DNA sequencing, we can more fully evaluate the importance of regulatory mechanisms, epigenetics, and CNVs upon selectable phenotypes.

ACKNOWLEDGEMENTS

Benjamin Greenbaum is the Eric and Wendy Schmidt Member of the Institute for Advanced Study and would like to gratefully acknowledge their support. Benjamin Greenbaum would like to thank Raul Rabadan and Bud Mishra for their reading and many helpful suggestions. He would also like to thank Gyan Bhanot, Vladimir Trifonov, and Hossein Khiabanian for many helpful discussions, and Lindsay Greenbaum for her careful reading of the manuscript.